Tugas Pekan ke-4 Semantik Vektor II

Batas pengumpulan: Jumat 9 Oktober 2020, pukul 10.59 pagi, melalui LMS

Deskripsi

Buatlah model word2vec Skip-Gram menggunakan library Gensim dari 100 artikel berbahasa Indonesia. Topik artikel bebas, akan lebih baik jika keseluruhan 100 dokumen tersebut topiknya serupa. Lakukan eksperimen dengan jumlah minimal kemunculan kata berbeda (misal di eksperimen pertama jumlah minimal kemunculan kata = 1, kemudian di eksperimen kedua jumlah minimal kemunculan kata = 5). Cara setting jumlah minimal kemunculan kata silakan cari di dokumentasi/tutorial Gensim. Panjang vektor yang akan dihasilkan silakan tentukan sendiri, boleh pakai nilai default (Panjang vektor/embedding = 100).

Setelah Anda membuat model word2vec Skip-Gram dengan 2 setting minimal kemunculan kata tersebut, lakukan eksperimen:

1. Akses representasi vektor/embedding sebuah kata dari 2 setting minimal kemunculan kata.
2. Hitung nilai *similarity* antar kata dengan memilih 3 pasang kata yang Anda perkirakan nilai similarity-nya (dari 2 model yang Anda peroleh dari 2 setting eksperimen):
   a) similarity > 0,5
   b) similarity > 0 dan similarity < 0,5
   c) similarity < -1 dan similarity > -0,5
   Nilai similarity yang Anda peroleh dari fungsi Gensim cantumkan di laporan, beri penjelasan apakah nilainya sesuai dengan yang Anda perkirakan di awal atau tidak. Jika tidak sesuai, berikan penjelasan dugaan sebabnya. Bandingkan hasil antara setting minimal kemunculan kata yang berbeda. [Petunjuk: silakan cari lebih lanjut makna similarity negatif pada vektor/embedding model word2vec].
3. Cari top-5 kata yang similar dengan sebuah kata tertentu, amati hasilnya, berikan analisis terhadap hasil yang Anda peroleh. Bandingkan hasil antara setting minimal kemunculan kata yang berbeda.
4. Lakukan visualisasi embedding dengan matplotlib. Amati apakah visualisasi yang diperoleh dapat memberikan informasi kata-kata yang mirip atau berkaitan maknanya. Bandingkan hasil antara setting minimal kemunculan kata yang berbeda.

Informasi yang harus dituliskan pada laporan:

1. Print-screen hasil eksperimen nomor 1.
2. Hasil eksperimen beserta penjelasan analisis eksperimen nomor 1-4.

File yang harus dikumpulkan:

1. Program dan kelengkapannya: 1 file kode program python (.py) + 100 file artikel (.txt) **dalam sebuah folder** + petunjuk menjalankan program (.txt).
2. Laporan: 1 file pdf, maksimum panjang laporan adalah 2 halaman.

Penilaian: 40% source code + 60% laporan

Detail penilaian:

a. Program:

- kebenaran implementasi pembacaan file masukan dari sebuah direktori/folder (10 poin)
- kebenaran implementasi pembangunan model dengan setting minimal kemunculan kata berbeda (10 poin)
- kebenaran pemanggilan fungsi similarity antar kata dan top-n similar (10 poin)
- ketepatan penggunaan fungsi visualisasi embedding (10 poin)

b. Laporan:

- kelengkapan (30 poin)
- analisis eksperimen poin 2-4 (30 poin)

Jika ada pertanyaan, silakan disampaikan melalui *channel* pekan_4_tugas_vektor_semantik_2 di slack.